BiMMuDa Documentation

December 2, 2022

1 Introduction

The BiMMuDa (Billboard Melodic Music Dataset) is a collection of MIDI files representing the main melodies of top *Billboard* year-end singles from 1950 to 2020. This paper describes the compilation, structure and features of the dataset in detail.

2 Motivation

The primary motivation for compiling the dataset was to improve statistical models of melodic cognition in Western listeners. Many important processes in music cognition are governed by statistical learning (Pearce, 2018). Listeners unconsciously construct and update internal models of music based on everyday exposure, in which the probability of a musical event or string of events is determined by the frequency at which the listener has heard it over their lifetime. This representation has been found to dictate listeners' expectations for melody (Pearce and Wiggins, 2006), harmony (Bigand et al., 2003) (Sears et al., 2019) and rhythm (Palmer and Krumhansl, 1990). Expectancy, in turn, is a central mechanism for several processes in music perception, including sequential segmentation (Saffran et al., 1999; Tillmann and McAdams, 2004), recognition memory (Agres et al., 2018; Bartlett and Dowling, 1980; Cuddy and Lyons, 1981), and emotion (Egermann et al., 2013; Gingras et al., 2016). Thus, much of our cognitive processing involving music can be explained with a single probabilistic model, which can be embodied computationally using statistical learning algorithms. For musical expectancy, the leading computational model is the Information Dynamics of Music (IDyOM), a variable-order Markov model (Pearce, 2005).

Since internal psychological models of music are derived from past exposure, in order for a statistical learning algorithm to correctly simulate an individual's musical expectations, its training data must encapsulate the individual's cumulative experience with music. To model the expectations of Western listeners, IDyOM is often trained on a collection of Canadian folk songs, German folk songs, and Bach chorales, and given IDyOM's success as a model of musical

expectancy, it seems to sufficiently represent Western musical structure. However, it can be argued that the data used to train models like IDyOM needs to be updated. For the vast majority of Western listeners, classical and folk music comprise only a small fraction of aggregate musical exposure. Overwhelmingly, the songs playing in homes, cars, coffee shops and dance clubs, the songs that Western listeners know by heart today, are from the popular genres of the twentieth and twenty first centuries (Sloboda and O'Neill, 2001). Models of music cognition may be improved if they are trained on data that reflects this reality, especially when modelling non-musicians, who are less likely to be familiar with classical or folk music (Sears et al., 2019). In addition, stimulus design for experiments in music cognition employs the classical and folk genres almost exclusively. While these stimuli elicit responses that are, to an extent, useful for understanding the faculties involved in music perception and cognition, one could argue that their relevance to these faculties' everyday functioning in most Western listeners is limited. Though it is impossible to replicate the contexts in which everyday listening occurs in an experimental setting, using stimuli Western participants are more likely to have heard in everyday life may yield more precise knowledge. The availability, quality, and size of datasets containing classical music compared to those with pop music is a major reason for their ubiquity in music research. Table 2 in Ji et al.'s recent survey on automatic music generation summarizes the known datasets. There are datasets of Western pop music, most notably the Lahk MIDI dataset (Raffel, 2016), a corpus of over 175,000 MIDI files, and the POP909 dataset (Wang* et al., 2020), which contains MIDI files and structural information for 909 popular songs. The largest datasets are suitable for model training, but quality within them is inconsistent, and their compilation appears opportunistic and quantity-driven. Popular songs that are likely to be familiar to many are heavily outnumbered by obscure music. The only dataset that may approximate the average Western listener's lifetime musical exposure is Mauch et al.'s dataset of *Billboard Hot 100* audio recordings from 1960 to 2010. However, there is no equivalent dataset in the symbolic domain, where many musical features are easier to extract. The BiMMuDa complements Mauch et al.'s dataset, as it provides MIDI representations for the main melodies of the most popular subset of the songs.

In summary, the BiMMuDa is a dataset both large enough to serve as a training set for statistical models in music cognition and of high enough quality to provide stimulus sets for experimental research in the field. It is intended to exemplify the average Western individual's lifetime exposure to music, especially when paired with Mauch et al.'s dataset, and document the history of Western pop music in a machine-readable way.

3 The Billboard Year-End Singles Charts

The dataset is intended to be a compilation of melodies from the five most popular songs of each year, from 1950 to 2020. The *Billboard* year-end singles chart, which is based on cumulative measures of single performance reported by

Billboard magazine, is the chosen measure of popularity for two reasons. Firstly, it is the only aggregate measure of popularity that existed as far back as 1950. Secondly, *Billboard* has adjusted their formula for determining chart rankings several times over its history to reflect the changing ways music is and can be enjoyed. In the 1950s they used record sales, airplay and jukebox activity data to assess performance, while today they incorporate data about online streams and paid digital downloads in addition to airplay data (Molanphy, 2013). This makes *Billboard*'s year-end chart a more robust indicator of popularity than any singular measurement of a song's performance.

However, *Billboard*'s charts are imperfect as measurements of popularity. The charts consulted for this dataset track year-end performance only in the United States, which may not be an adequate indicator of global performance or performance in other Western countries. Artists and record labels have found ways to manipulate the charts throughout the years (Andrews, 2018). *Billboard* policy has kept some extremely popular songs from ranking highly on the year-end charts or even appearing on them at all. Before 1991, due to the way chart positions were calculated, cumulative points for songs released towards the end of the year were often split over two year-ends, causing them to rank lower than they would have had the peak of their popularity occurred within a single year. In the 1990s, some songs achieved massive success but were not eligible for *Billboard*'s year-end singles chart because they were not available for purchase as singles; in 1998 the Goo Goo Dolls' "Iris" was number one for 18 weeks on *Billboard's Hot 100 Airplay* chart but did not appear on the year-end singles chart. These issues must be kept in mind, but *Billboard*'s singles charts are still considered the standard measures of a song's popularity, and the top five year-end singles from each year can provide insight about trends in popular Western music. The list of singles represented in the dataset can be found in the metadata (see Section 4.6).

4 Dataset Details and Compilation

The basic structure of the dataset is as follows: the top level contains folders for each of the 72 years. Each of these folders contains five more folders labelled 1-5, representing the top five singles from the relevant year. Within each of these folders are the MIDI files containing the melodies of the corresponding single[1].

It is worth explaining the process by which each song's melodies were compiled. There are two versions of the process; the version used for each song depended on whether or not a usable pre-existing MIDI file of the song was available, which was the case approximately 40 percent of the time. The process involves determining the long-term structure of the song and segmenting it

[1]There are exceptions to this: for some songs it was determined that no main melody existed, and thus the folders corresponding to those songs are empty. Conversely, some year-end top singles required two folders because they are double-sided; see the Appendix for more information.

appropriately, identifying the main melodies, either transcribing the melodies or extracting them from the existing MIDI, concatenating the melodies together according to the song's long-term structure, and recording the song's metadata. The details of each step are below.

4.1 Segmentation

Most pop songs consist of sections, such as the verse, chorus, and bridge, repeated according to some long-term structure (e.g., AABA, ABABABCBB). The melodies of each section are stored in separate files, so it was best to identify the sections and their arrangement within the song before performing any transcription.

Almost every song that ranks highly on a *Billboard* year-end chart can easily be found on Spotify. I listened to each song and noted its long-term structure, which is a straightforward endeavor the vast majority of the time. Lyrics websites often correctly segment the song's lyrics according to its long-term structure, and section boundaries are usually signaled with very noticeable changes in timbre. Occasionally, segmentation was subjective: sometimes it is unclear whether one section differs enough from another to be considered unique, and if two short sections always occur together but are very distinct in some way (e.g., they feature different sets of instruments), it must be decided whether to deem them separate sections or combine them. However, with this dataset, segmentation difficulties were minimal.

4.2 Main Melody Identification

For almost all songs, either the entirety of the main melody is sung by a lead vocalist, or a lead vocalist takes turns with one or more featured vocalist. There are a few exceptions: some top singles are orchestral pieces where the main melody is played by one or more instruments and often switches between instruments. For some songs there was no main melody and therefore no melody transcription/extraction was performed. See the Appendix for more information.

If a multi-track MIDI file for the song was available, main melody selection simply meant picking the track (or tracks) corresponding to the main melody of the song and segmenting it into the parts identified in the previous step. If there was no usable pre-existing MIDI, main melody identification and transcription occurred simultaneously.

Main melody selection can be a source of error for a few reasons. Sometimes there are two equally salient melodies playing simultaneously, e.g., duets, overlapping melodies. The dataset is strictly monophonic, so one melody was selected as the main one, sometimes quite arbitrarily. Often the end of the main melody of one section will overlap with the beginning of the main melody of the next section, and in these cases the melodies had to be clipped to remove overlap. Finally, vocal harmonies can be very close in volume to the main melody, making the main melody difficult to isolate.

4.3 (With Pre-Existing MIDI) Melody Correction

There are many websites that provide free MIDI files for pop songs, which were searched thoroughly to minimize the amount of manual transcription that would be required to complete the dataset. The following websites were searched:

- freemidi.org
- bitmidi.com
- midiworld.com
- mididb.com
- musescore.com

MIDI files were found for about 90% of songs, but less than half were usable for this project. Many were karaoke MIDIs, so the main melody was not present. Others were type 0 MIDI files in which all voices resided on a single track. It is usually easier to manually transcribe melodies than untangle the main melody from such files.

In the highest quality MIDI files, each instrument or voice resides has its own track, so the track corresponding to the lead vocal/main melody can be easily isolated. If this is the case, after main melody detection it is only necessary to thoroughly check the melodies by listening closely to the song, comparing the audio with the transcribed melody, and making corrections as needed.

4.4 (Without Pre-Existing MIDI) Melody Transcription

If no high-quality MIDI file for the song was available, I manually transcribed the main melodies through repeatedly listening of the song bar-by-bar. All MIDIs were segmented, transcribed and checked in FL Studio 20. For every melody, the instrument used is the "FL Keys" stock plugin on its default setting. The tempo, meter and key of each MIDI is the same as those of the song it represents, as estimated by Tunebat.

When transcribing (or making corrections), my primary focuses were pitch and onset time. All onsets are quantized after transcription/correction with `music21`'s `quantize` function. I did not examine velocity (volume); all the notes in manually transcribed melodies have the same velocity, and I left the velocities in pre-existing MIDIs untouched. Inevitably, I made errors when transcribing and correcting melodies.

After transcription/correction, the melodies of each section are exported into separate MIDI files. The files are named according to the year, rank, and order in which the section appears in the song. For example, "196501_3.mid" contains the main melody of the third section of the number one song in 1965, according to *Billboard*.

A few songs will have MIDI files with a "misc" suffix. These are usually hooks in the instrumental that seemed to greatly contribute to the "catchiness"

of the song but may not be considered part of the main melody. See, for example, "201704_misc.mid", which contains the top voice of the piano in "Humble" by Kendrick Lamar. These kinds of melodies are not included in the concatenated files (see the next subsection).

4.5 Concatentation

The melodies from the different sections of the song are also pasted together according to the long-term structure to roughly recreate the full song. These kinds of MIDI file are named with the "full" suffix. The length of these kinds of files are usually *not* equal to the length of the song audio, since they do not accurately represent gaps between sections (where there is no main melody). Despite this, hopefully these files will aid in the study of long-term structure in Western pop music. Depending on how the dataset is being used, it might be best to exclude these files from analysis, since they are in a sense redundant.

Feature	**Description**
Title	title of the song
Artist	artist(s), including any featured artists
Year	year in which the song appeared in the top five of the *Billboard* year-end singles chart
Position	song's position on the *Billboard* year-end singles chart
BPM	tempo of the song in beats per minute, as estimated by Tunebat
Link to Audio	Spotify or YouTube link to the song
Tonic	tonic of the song, as estimated by Tunebat
Mode	mode (major/minor), as estimated by Tunebat
Number of Parts	number of melodies in the song, excluding "misc" melodies
Number of Words	number of words in the lyrics file, including repeated words and sections
Number of Syllables	number of syllables in the lyrics file, using the algorithm in (syl, 2013)

Table 1: BiMMuDa per-song feature descriptions

4.6 Lyrics and Metadata

The lyrics of each song were obtained from Genius.com and saved in a .txt file, if applicable. See the Appendix for a list of songs with no lyrics. Finally, metadata is available via two .csv files, one for per-song features and one for per-melody features, described in Tables 1 and 2. Summary statistics are given in the next subsection.

Feature	Description
Melody ID	name of the file (e.g., "196001_1")
Length (seconds)	length of the MIDI file in seconds
Length (notes)	number of note events in the melody
Tonality	conformance to one of twelve keys in Western music (Krumhansl, 1990)
Markov Model Information Content (MMIC)	information-theoretic unpredictability
Contour Score	average distance between consecutive pitches, in MIDI pitch note numbers
Onset Density	average number of notes per second
Normalized Pairwise Variability Index (nPVI)	contrast between consecutive onsets (Patel and Daniele, 2003)
Syncopation Score	degree of onset deviation from metrical accents

Table 2: BiMMuDa per-melody feature descriptions. Italicized features are described in more detail in the Supplementary Index of (cite PNAS paper)

5 Summary Statistics

5.1 Per-Song Features

Table 3 summarizes the numeric per-song features, and their correlation matrix is visualized in Figure 1. The mean number of parts (melodies) per song, overall and per decade, is given in Figure 2. The distributions of the Mode, Tonic, and Key features are visualized in Figures 3, 4, and 5, respectively. Finally, the BPM feature is visualized per decade in Figure 6.

Feature	Mean	Median	Std. Dev	Range
Number of Parts	3.011	3.0	1.12	0 - 7
BPM	106.0	104.0	24.51	57 - 174
Number of Words	327.8	296.0	176.4	0 - 896
Number of Syllables	403.7	368.0	217.0	0 - 1064

Table 3: Summary Statistics for BiMMuDa per-song features

Overall, a song has 3 melodies on average, and about two-thirds of the songs have 2-4 melodies. The major mode is more common in every decade, though the major/minor split becomes more even after the 1960s. The distribution of tonics is relatively even, with songs in C and G being particularly popular in the 1960s. Mean BPM cycles over time: faster songs (BPM = 115-129 and above) are more common in the 1960s and post-2000, while the prevalence of slower songs (in particular, with $57 \leq BPM \leq 69$ and $85 \leq BPM \leq 99$) peaks in the 1990s. Per-song features are not strongly correlated, with the exception of Number of Words and Number of Syllables.

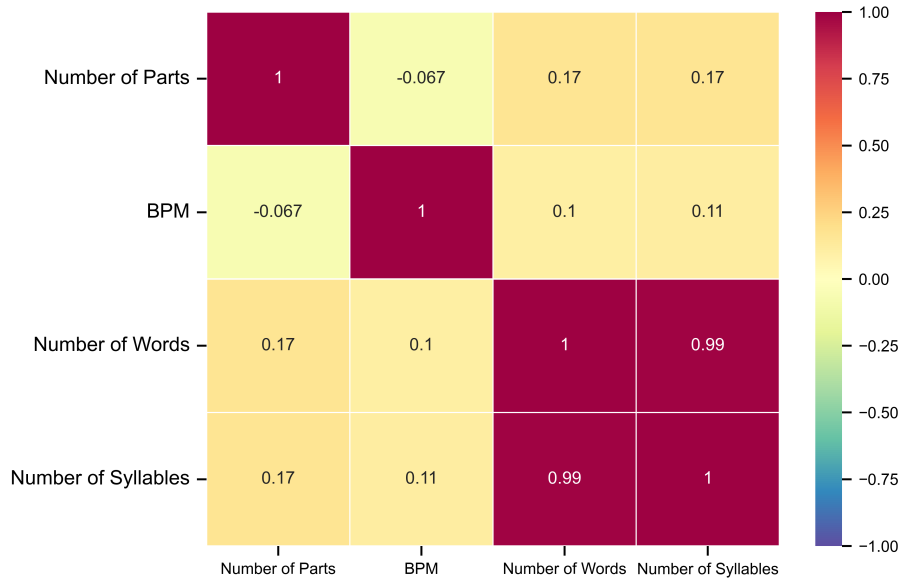

Figure 1: Correlation Matrix for Per-Song Features

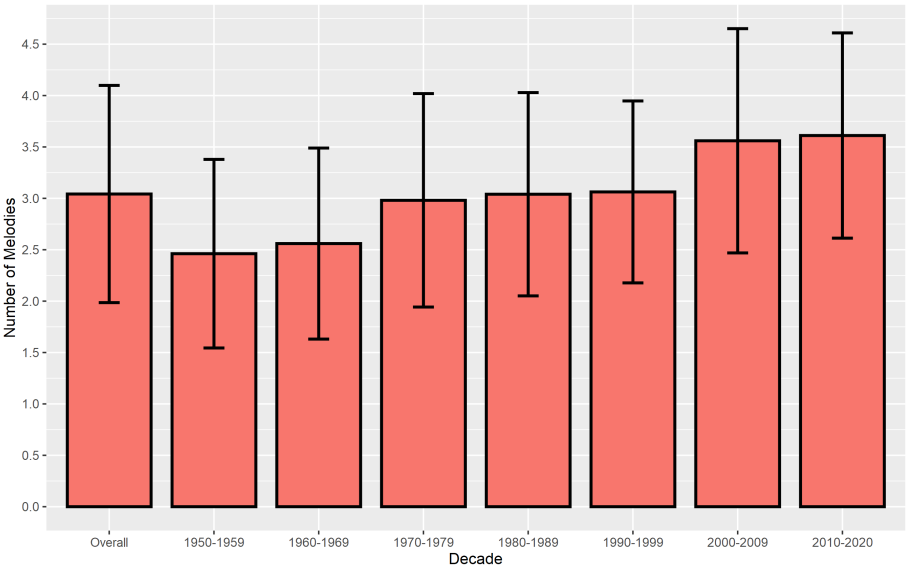

Figure 2: Mean number of melodies per song, over the entire dataset and by decade, with error bars

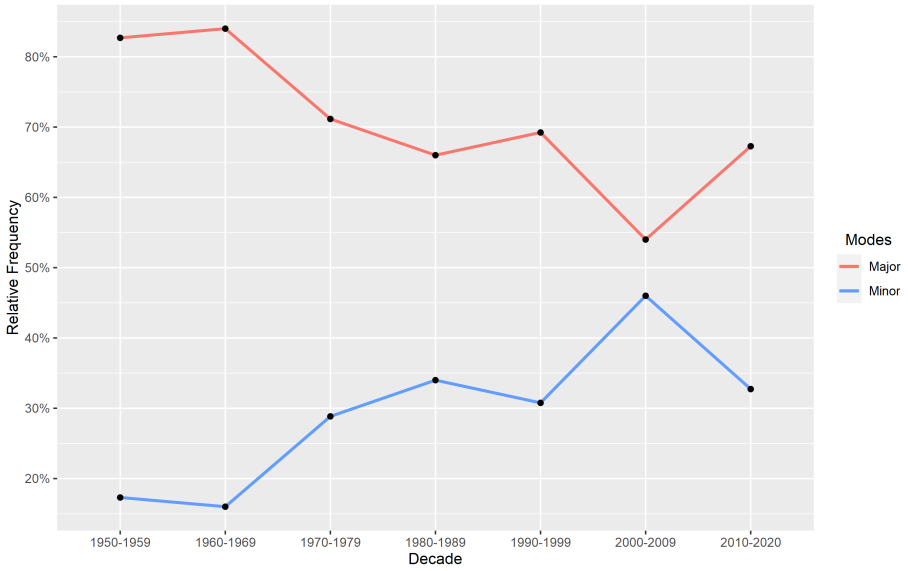

Figure 3: Frequency of major and minor modes by decade

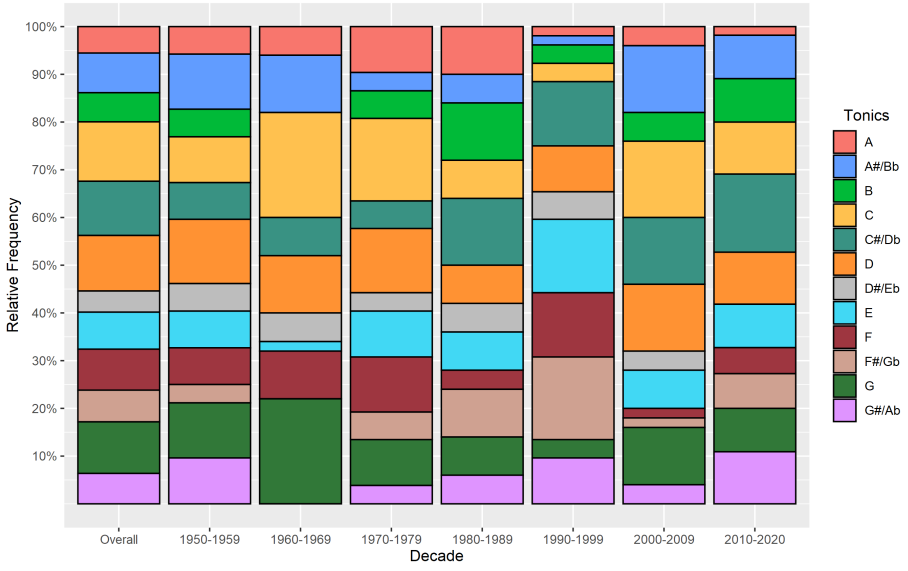

Figure 4: Distribution of tonics, overall and by decade

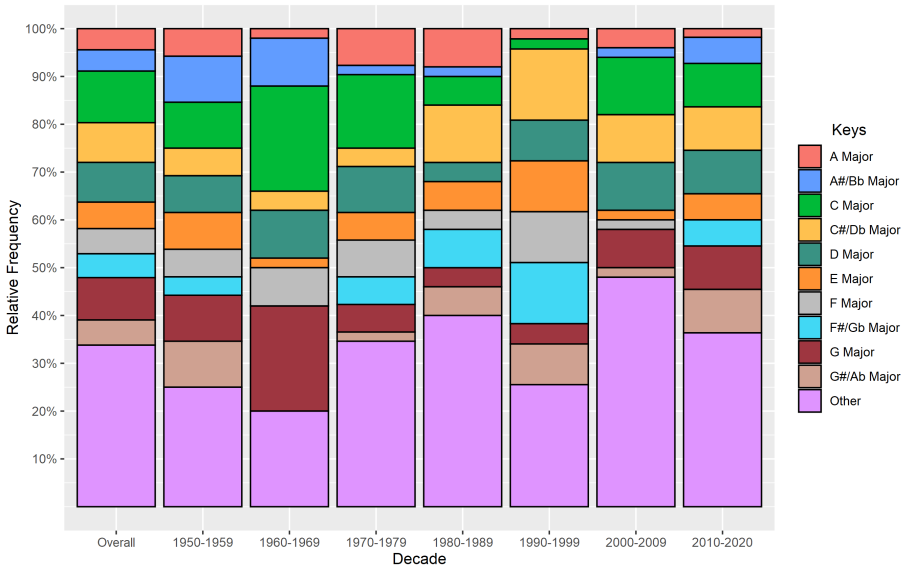

Figure 5: Distribution of the 10 most common keys, overall and by decade

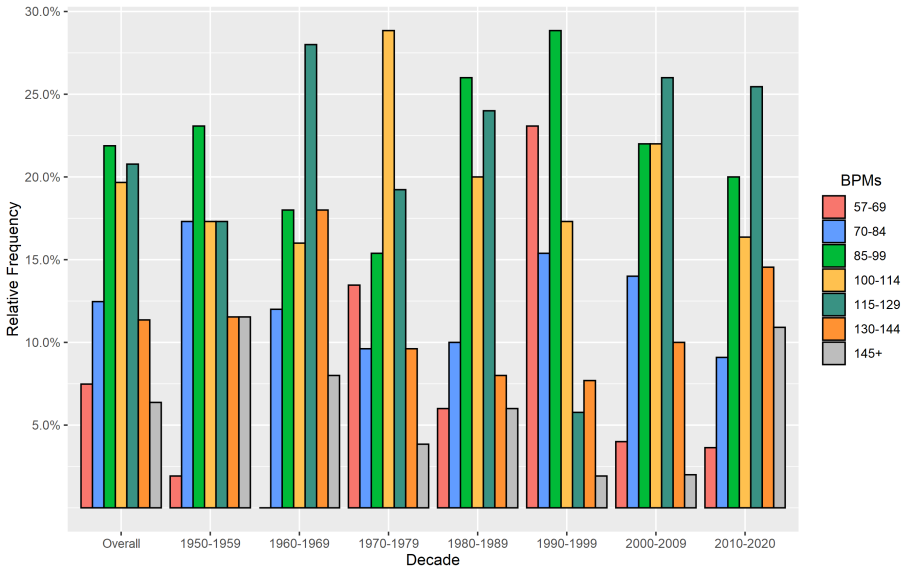

Figure 6: Distribution of BPMs, overall and by decade

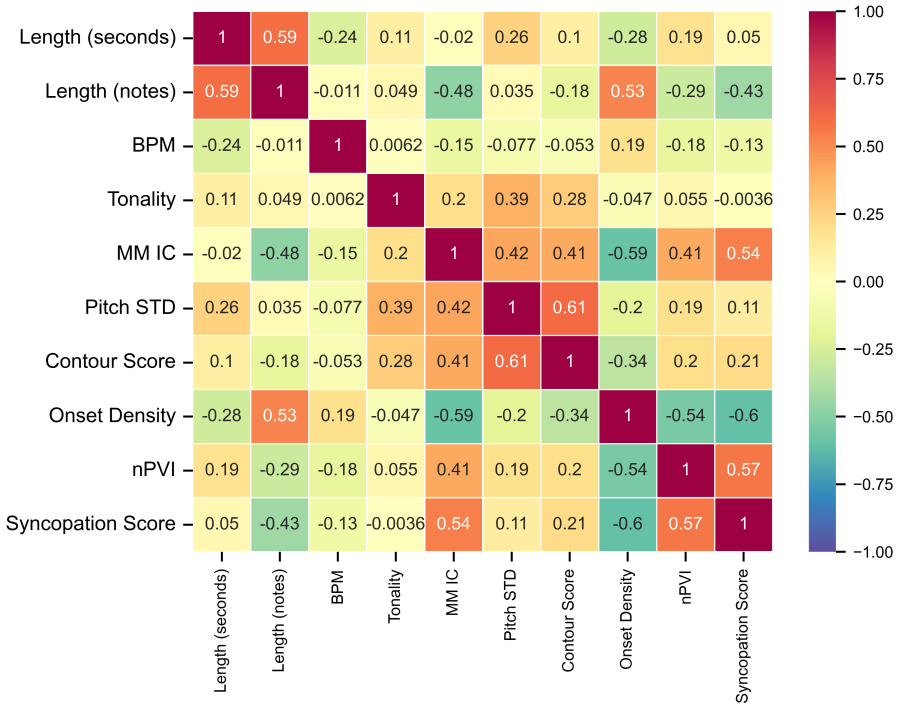

Figure 7: Correlation Matrix for Per-Song Features

5.2 Per-Melody Features

Table 4 and Figure 7 summarize BiMMuDa's per-melody features and their correlations, respectively. Figure 8 visualizes the basic per-melody features over time. The remaining per-melody features are analyzed in detail in (CITE PNAS PAPER).

Feature	**Mean**	**Median**	**Std. Dev**	**Range**
Length (seconds)	22.49	20.53	9.521	3.010 - 65.26
Length (notes)	47.76	44.0	23.36	4 - 169
BPM (per melody)	105.4	104.0	23.94	57 - 174
Tonality	0.736	0.744	0.105	0.402 - 0.982
MM IC	3.595	3.563	1.003	0.285 - 6.123
Pitch STD	3.002	2.877	1.138	0.0 - 9.719
Contour Score	2.111	2.0968	0.829	0.0 - 10.85
Onset Density	2.223	2.114	0.836	0.443 - 5.690
nPVI	41.15	39.56	20.58	0.0 - 128.7
Syncopation Score	2.305	2.200	0.804	0.233 - 6.882

Table 4: Summary Statistics for BiMMuDa per-melody features

Decade	Number of Melodies	Number of Note Events	Sum of Melody Lengths (minutes)
1950-1959	128	5293	54.55
1960-1969	128	4854	47.47
1970-1979	155	6741	59.79
1980-1989	152	6421	53.83
1990-1999	147	6890	58.16
2000-2009	**178**	**10717**	**66.82**
2010-2020	**195**	**11239**	**68.11**

Table 5: Number of melodies, note events, and minutes in the dataset, by decade

While melody length in seconds stays relatively consistent over time, melody length in note events increases significantly in the 1990s. Additionally, as shown in Figure 2, post-2000 songs have more melodies. This uncovers an important point: BiMMuDa is unbalanced in favor of the two most recent decades, especially in terms of note events. See Table 5. 2000-2020 covers about 30% of the represented years, it has 34% of the melodies, 42% of the total note events, and 33% of the total "amount" of melody, in seconds. See Table 5.

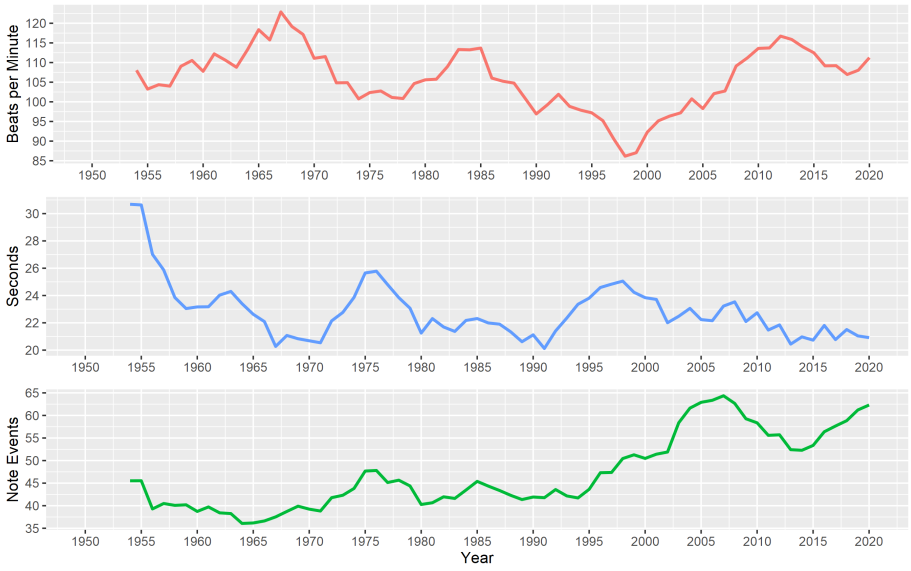

Figure 8: Five-year rolling mean over time for Length (seconds), Length (notes) and BPM

A Appendix

A.1 Top Singles with No Main Melody

The following high-ranking singles were deemed to have no main melody, though all have some melodic aspect. The folders associated with these singles are empty, except for 201704, the folder for "Humble" by Kendrick Lamar, which has one "misc" file.

- "Baby Got Back" by Sir Mix-a-Lot (199202)
- "Jump" by Kriss Kross (199203)
- "Whoomp! There It Is" by Tag Team (199302)
- "Can't Nobody Hold Me Down" by Diddy feat. Mase (199705)
- "Humble" by Kendrick Lamar (201704)

A.2 Top Singles with Non-Vocal Melodies

Almost all melodies in the dataset are vocal melodies, with the exception of the following 13 singles. The folders associated with these songs will have no lyrics file.

- "Third Man Theme" by Anton Karas (195003)

- "Blue Tango" by Leroy Anderson (195201)
- "The Song from Moulin Rouge" by Percy Faith (195301)
- "Cherry Pink and Apple Blossom White" by Perez Prado (195501)
- "Autumn Leaves" by Roger Williams (195504)
- "Lisbon Antigua" by Nelson Riddle (195603)
- "Patricia" by Perez Prado (195805)
- "The Theme from 'A Summer Place'" by Percy Faith (196001)
- "Stranger on the Shore" by Acker Bilk (196201)
- "The Stripper" by David Rose (196205)
- "Love is Blue" by Paul Mauriat (196802)
- "Love's Theme" by Love Unlimited Orchestra (197403)
- "Harlem Shake" by Baauer (201304)

A.3 Double-Sided Singles

For double-sided singles, *Billboard* used to list both sides on the charts with the best-performing side listed first, regardless of how the less popular side performed. Then in 1969, *Billboard* began listing the A and B-sides together only if both sides received significant airplay. There are six double-sided singles in the dataset:

- "All I Have to Do Is Dream / Claudette" by the Everly Brothers (195802)
- "Don't / I Beg of You" by Elvis Presley (195803)
- "Maggie May / Reason to Believe" by Rod Stewart (197102)
- "I Feel the Earth Move / It's Too Late" by Carole King (197103)
- "Something About The Way You Look Tonight / Candle in the Wind" by Elton John (199701)
- "Foolish Games / You Were Meant for Me" by Jewel (199702)

Here, the melodies in the A and B-sides are stored in separate folders, distinguished by "a" and "b" suffixes ("195803a" and "195803b", etc.)

References

Counting syllables in the english language using python, 2013. URL https://eayd.in/?p=232.

K. Agres, S. Abdallah, and M. Pearce. Information-theoretic properties of auditory sequences dynamically influence expectation and memory. *Cognitive Science*, 42:43–76, 2018.

T. Andrews. Billboard's charts used to be our barometer for music success. are they meaningless in the streaming age?, 2018. URL https://www.washingtonpost.com/news/arts-and-entertainment/wp/2018/07/05/billboards-charts

J. Bartlett and W. Dowling. Recognition of transposed melodies: a key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, 6:501–515, 1980.

E. Bigand, B. Poulin, B. Tillmann, F. Madurell, and D. D'Adamo. Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology: Human perception and performance*, 29(1):159–171, 2003.

L. Cuddy and H. Lyons. Musical pattern recognition: a comparison of listening to and studying tonal structures and tonal ambiguities. *Psychomusicology*, 1: 15–33, 1981.

H. Egermann, M. Pearce, G. Wiggins, and S. McAdams. Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, and Behavioral Neuroscience*, 13: 533–553, 2013.

B. Gingras, M. Pearce, M. Goodchild, R. Dean, G. Wiggins, , and S. McAdams. Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, 42(4):594–609, 2016.

S. Ji, J. Luo, and X. Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. 2020.

C. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, 1990.

M. Mauch, R. MacCallum, M. Levy, and A. Leroi. The evolution of popular music: Usa 1960-2010. *R. Soc. opensci.*, 2(5), 2015.

C. Molanphy. How The Hot 100 Became America's Hit Barometer, 2013. URL https://www.npr.org/sections/therecord/2013/08/16/207879695/how-the-hot-100-became-america

C. Palmer and C. Krumhansl. Mental representation for musical meter. *Journal of Experimental Psychology: Human perception and Performance*, 16:728–741, 1990.

Aniruddh D Patel and Joseph R Daniele. An empirical comparison of rhythm in language and music. *Cognition*, 87(1):B35–B45, 2003.

M. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, School of Informatics, City University, London, 2005.

M. Pearce. Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, 1423:378–395, 2018. doi: 10.1111/nyas.13654.

M. Pearce and G. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405, 2006.

C. Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, 2016.

J. Saffran, E. Johnson, R. Aslin, and E. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.

D. Sears, M. Pearce, J. Spitzer, W. Caplin, and S. McAdams. Expectations for tonal cadences: Sensory and cognitive priming effects. *Quarterly Journal of Experimental Psychology*, 72:1422–1438, 2019.

J. Sloboda and S. O'Neill. Emotions in everyday listening to music. In P. Juslin and J. Sloboda, editors, *Series in affective science*, pages 415–429. Oxford University Press, Oxford, 2001.

B. Tillmann and S. McAdams. Implicit learning of musical timbre sequences: Statistical regularities confronted with acoustic (dis)similarities. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(5):1131–1142, 2004.

Ziyu Wang*, Ke Chen*, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Guxian Bin, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. In *Proceedings of 21st International Conference on Music Information Retrieval, ISMIR*, 2020.